AL-TP-1991-0001

AD-A237 484

ARMSTRONG LABORATORY

SHORT-TERM TEST-RETEST RELIABILITY OF AN EXPERIMENTAL VERSION OF THE BASIC ATTRIBUTES TEST BATTERY

Thomas R. Carretta

HUMAN RESOURCES DIRECTORATE
MANPOWER AND PERSONNEL DIVISION
Brooks Air Force Base, TX 78235-5000

June 1991

Interim Technical Paper for Period March 1989 - February 1991

91-03607

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235-5000

NOTICES

This paper has been reviewed and is approved for publication.

Thomas R Carretta

THOMAS R. CARRETTA
Project Scientist

WILLIAM E. ALLEY, Technical Director
Manpower and Personnel Division

MICHAEL W. BIRDLEBOUGH, Colonel, USAF
Chief, Manpower and Personnel Division

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (Leave blank)

2. REPORT DATE
June 1991

3. REPORT TYPE AND DATES COVERED
Interim Technical Paper – March 1989 – February 1991

4. TITLE AND SUBTITLE
Short-Term Test-Retest Reliability of an Experimental Version of the Basic Attributes Test Battery

5. FUNDING NUMBERS
PE – 62703
PR – 7719
TA – 18
WU – 45

6. AUTHOR(S)
Thomas R. Carretta

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Armstrong Laboratory
Human Resources Directorate
Manpower and Personnel Division
Brooks Air Force Base, Texas 78235-5000

8. PERFORMING ORGANIZATION REPORT NUMBER
AL-TP-1991-0001

9. SPONSORING/MONITORING AGENCY NAMES(S) AND ADDRESS(ES)

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution is unlimited.

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 words)

Two hundred forty-seven (247) USAF pilot candidates commissioned through the Reserve Officer Training Corps (ROTC) were tested on an experimental form of the Basic Attributes Test (BAT) battery twice (once on 2 consecutive days) at the beginning of a pre-Undergraduate Pilot Training (UPT) Flight Screening Program. The purpose of this investigation was to examine the short-term test-retest reliability of the BAT battery. Results indicate that on the second test administration 69.6% of the subjects improved their overall performance. There was some evidence of regression toward the mean on the second administration. Only 66 out of 124 subjects (53.2%) who had BAT composite scores above the median on the first administration improved their scores on the second administration. However, 106 out of 123 subjects (86.2%) who scored below the median on the first administration improved their scores on the second administration. Despite this, there was a modest correlation between subjects' first and second administration BAT composite scores (Pearson $r = .56$; Spearman $r = .55$). These results are consistent with a study conducted by the US Navy that examined test-retest reliability for a battery of cognitive speed tests (Saccuzzo & Larson, 1987). The test-retest correlations in the Saccuzzo and Larson study were of somewhat lower magnitude than in the present study. However, the test-retest interval was 10 days in the US Navy study versus 1 day in the present study.

(Continued)

14. SUBJECT TERMS
computerized testing
pilot candidate selection
reliability
undergraduate pilot training

15. NUMBER OF PAGES
28

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2-89)
Prescribed by ANSI Std. Z39-18
298-102

13. Abstract (Concluded)

The magnitude of the test-retest correlations in the present investigation may be underestimated because of reduced test length and preselection of subjects on the basis of the AFOQT. The stability of the pilot candidate selection composite could be improved by removing from the composite those test scores with low reliability. Implications for a BAT test-retest policy and for a forthcoming measurement and metric equivalency study are discussed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

This project was conducted under work unit 77191845 in support of Request for Personnel Research (RPR) 78-11, Selection for Undergraduate Pilot Training, issued by Air Training Command.

Appreciation is extended to Mr Roy Chollman and his staff for support in scheduling and testing subjects, to Mr Bob Levine from OAO, Incorporated for his assistance in preparing the data files and programming the data analyses, and to Mr Robert Picacio for administrative support. The author also extends thanks to Maj Dave Perry, Dr Rick Siem, Dr Laurie Walters, Dr Joseph L. Weeks, Dr William E. Alley, and Dr Malcolm Ree for their comments and technical support during this project.

SHORT-TERM TEST-RETEST RELIABILITY OF AN EXPERIMENTAL VERSION OF THE BASIC ATTRIBUTES TEST BATTERY

SUMMARY

Two hundred forty-seven (247) USAF pilot candidates commissioned through the Reserve Officer Training Corps (ROTC) were tested on an experimental form of the Basic Attributes Test (BAT) battery twice (once on 2 consecutive days) at the beginning of a pre-UPT Flight Screening Program. The purpose was to examine the short-term test-retest reliability of the BAT battery. Results for the second test administration indicate that 69.6% of the subjects improved their overall performance; also there was some evidence of regression toward the mean. For subjects who scored above the median BAT composite score on the first administration, only 66 of 124 (53.2%) improved their scores on the second administration. For those who scored below the median on the first administration, 106 of 123 (86.2%) improved their scores on the second administration. Despite this, there was a modest correlation between subjects' first and second administration BAT composites (Pearson r = .56; Spearman r = .55).

These results are consistent with previous research regarding the test-retest reliability of computer-administered cognitive abilities tests (Roznowski, 1989; Saccuzzo & Larson, 1987). The Roznowski (1989) study examined test-retest reliability for a broad cognitive abilities test battery, whereas Saccuzzo and Larson (1987) used an experimental US Navy battery of cognitive speed tests. The magnitude of the test-retest correlations in the Roznowski and the Saccuzzo and Larson studies were somewhat lower than in the present study. However, the test-retest interval was 2 weeks in the Roznowski study, 10 days in the US Navy study, and only 1 day in the present investigation.

The magnitude of the test-retest correlations found here may be underestimated because of reduced test length and preselection of subjects on the basis of the AFOQT. Improvement in the stability of the pilot candidate selection composite could be made by removing test scores with low reliability from the composite. Implications for a BAT test-retest policy and for a future measurement and metric equivalency study are discussed.

I. INTRODUCTION

Background

Between February 1942 and July 1955, measurement of perceptual and psychomotor abilities was an important component of the United States Air Force (USAF) pilot candidate selection procedure. During that time, several testing devices provided reliable measures of perceptual and motor skills that were useful for classifying aircrew applicants into job specialties (pilot versus navigator) and for predicting preliminary flying training performance (Passey & McLaurin, 1966). Apparatus-based testing was discontinued primarily for administrative reasons including the decision to decentralize the testing process and the difficulty in keeping the electro-mechanical testing devices calibrated and the test administration procedures consistent across test sites.

Since 1955, the variables considered in USAF pilot candidate selection have included medical fitness, academic performance, aptitude test scores (e.g., Air Force Officer Qualifying Test [AFOQT]), background/biographical data (e.g., type of college degree, age), and previous flying experience. Although this set of variables has demonstrated a consistent relationship to Undergraduate Pilot Training (UPT) performance, a desire to improve the pilot candidate selection process, as well as recent improvements in computer technology, has generated a renewed interest in apparatus testing as an additional selection variable.

At the request of the USAF Air Training Command (ATC), we began a multi-year research and development (R&D) program designed to improve selection procedures for USAF pilot training candidates. Two important research objectives were as follows: (a) to take advantage of state-of-the-art computerized testing technologies for the development of accurate and reliable measures of perceptual and motor skills, and (b) to investigate the utility of these test measures for improving current pilot candidate selection procedures (Bordelon & Kantor, 1986).

As part of this R&D effort, a series of studies were conducted using a computerized test device known as the Basic Attributes Test (BAT) system. Results showed that individual differences in hand-eye coordination, information processing ability, personality, and attitudes are useful for improving prediction of UPT performance beyond that provided by currently used procedures (Bordelon & Kantor, 1986; Carretta, 1989, 1990b; Kantor & Carretta, 1988). As a consequence, USAF personnel managers have decided to operationally implement apparatus test scores for use as an aid, in combination with other information, to USAF pilot candidate selection and classification decisions.

Research Objective

Prior to full-scale operational implementation of apparatus testing of USAF pilot applicants, several research issues need to be addressed. The primary research objective of the present investigation was to evaluate short-term test-retest score profiles for the Basic Attributes Test (BAT) battery. In an operational test environment, some pilot applicants may be permitted to retest after a fairly short time interval (e.g., test administration problems resulting in a need for a retest; medical problems during the test). If large short-term retest gains occur, it may be necessary to adjust test scores on the retest. Another research issue involves evaluating equivalency between the current BAT system and software and a second-generation operational test device under development. Full-scale testing of USAF pilot applicants cannot be supported by the current experimental BAT prototypes because computerized testing devices are too few in number and are expensive to acquire and maintain. When the first production prototypes become available, a study will be performed to evaluate measurement and metric equivalency of the current and new test systems. Measurement equivalency is achieved if the same tests programmed on both machines measure the same constructs (i.e., construct validity) and produce the same rank-ordering of subjects. Once measurement equivalency has been achieved, the current and new systems can be tested for metric equivalency. Metric equivalency exists if scores from the two machines have the same raw score distributions (i.e., means, standard deviations, skew, kurtosis).

If measurement equivalency is achieved, the new machines can be accepted as replacements for the current machines. If metric equivalency is achieved, the pilot candidate selection equations developed on the experimental BAT system can be used with scores obtained from the operational system.

Purpose

The primary purpose of the present investigation was to evaluate BAT score profiles and reliability over a short-term retest. Results may be used to help formulate a short-term BAT retest policy (i.e., under what conditions, if any, to allow a short-term retest; whether or not to adjust test scores on a short-term retest).

Results from this study also may provide some guidance regarding the amount of agreement that can be expected between scores from the experimental BAT system and those from the

BAT systems currently under development. It should be noted that some reliability estimates may be underestimated, as it was desirable to use shortened versions of some tests (Encoding Speed, Mental Rotation, Item Recognition, Time-Sharing, and Activities Interest Inventory) to accommodate time constraints and to minimize respondent fatigue effects. Previous reliability estimates for full-length versions of these tests suggested that they could be shortened without seriously reducing their internal consistency (Carretta [1990a] or see Table 5).

Before measurement equivalency between the experimental and operational test systems can be evaluated, a standard of reliability must be established. Because system reliability establishes an upper limit on the amount of score agreement that can be expected between experimental and operational scores, it is necessary to evaluate test-retest relationships for the experimental BAT system.

To date, test-retest relationships have been examined for only the two psychomotor coordination tests from the experimental BAT system (Two-Hand Coordination and Complex Coordination; Mercatante, 1988). Two hundred thirty-three (233) USAF pilot candidates were tested at the beginning of a light aircraft screening program before any training had taken place. They were retested 2 to 3 weeks after completing the program which involved about 14 hours of flying in a two-seat, single-engine aircraft. Test-retest reliability coefficients for the test scores ranged from .16 to .46. Mercatante (1988) suggests that the low test-retest reliability estimates may have been the result of the flying training that occurred between the two test administrations (i.e., flying experience may have affected psychomotor performance).

A study conducted by the US Navy regarding test-retest reliability of a cognitive speed test battery reported reliability coefficents that ranged from .01 to .66 (Saccuzzo & Larson, 1987). In another study, test-retest reliability coefficients between .23 and .87 were reported for a different computerized cognitive abilities test battery (Roznowski, 1989). The test-retest interval was 10 days in the US Navy study and 2 weeks in the Roznowski (1989) study.

II. METHOD

Subjects

The subjects in the present investigation were 247 USAF pilot training candidates who were commissioned through the Air Force Reserve Officer Training Corps (AFROTC). These students already had been chosen for pilot training, in part on the basis of their Air Force Officer Qualifying Test (AFOQT) scores.

Instrumentation

AFOQT. The AFOQT is a paper-and-pencil aptitude test battery used to select civilian or prior-service applicants for officer precommissioning training programs and to classify commissioned officers into aircrew job specialties (pilot vs. navigator training). The battery consists of 16 subtests that assess five ability domains: verbal, quantitative, spatial, aircrew interests/aptitude, and perceptual speed (Skinner & Ree, 1987). Fourteen of the 16 AFOQT subtests are used to compute the Pilot and Navigator-Technical composite scores used in the operational selection of pilot candidates (US Air Force, 1983).

Basic Attributes Test Battery. The BAT battery used in this study consisted of eight computerized tests that assessed individual differences in psychomotor coordination, information processing ability, personality, and attitudes. The types of scores generated from these tests include tracking error, response time, response accuracy, and response choice. Table 1 provides

a brief summary of this battery. A more detailed description is provided in the Appendix. To keep test administration time to a minimum, it was necessary to use shortened versions of five of the eight tests.

Table 1. **Basic Attributes Test (BAT) Battery Summary**

Test Name	Length (mins)	Attributes Measured	Types of Scores
Test Battery Introduction	10	Biographical Data	Age, gender, flying experience
Two-Hand Coordination (Rotary pursuit)	10	Tracking & Time-Sharing Ability in Pursuit	Tracking error
Complex Coordination (stick and rudder)	10	Compensatory Tracking Involving Multiple Axes	Tracking error
Encoding Speed 2	15	Verbal Classification	Response time, response accuracy
Mental Rotation 2	15	Spatial Transformation	Response time, response accuracy
Item Recognition 2	10	Short-Term Memory, Storage, Search and Comparison	Response time, response accuracy
Time-Sharing 3	20	Higher-Order Tracking Ability, Learning Rate & Time-Sharing	Tracking difficulty, response time, dual-task performance
Self-Crediting Word Knowledge	10	Self-Assessment Ability, Self-Confidence	Response time, response accuracy, bet
Activities Interest Inventory 2	5	Survival Attitudes	Response time, number of high-risk choices

Note. The Encoding Speed 2, Mental Rotation 2, Item Recognition 2, Time-Sharing 3 and Activities Interest Inventory 2 tests used in this study are shortened versions of these tests. The shortened versions were used due to time constraints that forced a reduction in the length of the BAT battery.

Apparatus

The BAT apparatus consists of a microcomputer and monitor built into a ruggedized chassis with a glare shield and side panels designed to minimize distractions. The subjects responded to the tests by manipulating (individually or in combination) a dual-axis joystick on the right side, a single-axis joystick on the left side, and a keypad in the center of the test unit. The keypad included keys labled O to 9, an ENABLE key in the center, and a bottom row with YES and NO keys and two other keys labled S/D for same/different responses and L/R for left/right responses.

Procedure

All subjects were enrolled in a 4-year college program. They were tested on the AFOQT either prior to entering college or while an undergraduate; they were tested on the BAT battery while attending a pre-UPT light aircraft screening program conducted in the summer following their junior year in college. Subjects were informed that the investigation involved evaluating experimental tests being considered for operational use. They also were told their performance would not affect their status in the program, would be kept confidential, and would be used only for research purposes.

Subjects were tested on the BAT battery at the beginning of the Flight Screening Program (FSP) prior to receiving any flying time. Each subject was tested twice on the BAT battery (once on 2 consecutive days).

The BAT battery used in this study required about 2 hours to complete, including programmed breaks between the tests. After the test administrator briefed the subjects and initialized the test battery, the test session was self-paced by each subject. Programmed breaks of 1 or 2 minutes between tests were included in order to reduce mental and physical fatigue.

Approach

Prior to establishing a short-term BAT retest policy or performing the equivalency study between the experimental BAT system and the operational BAT system, it is necessary to evaluate the stability of proposed tests administered on the experimental test device. Test-retest reliability estimates on the experimental BAT system provide an upper limit on the level of agreement that can occur between the experimental and operational systems.

Internal consistency reliability (i.e., Cronbach's alpha [Cronbach, 1951]) was estimated for each BAT score for both test administrations. Test-retest reliability (i.e., correlation between forms and Guttman split-half [Guttman, 1945]) were utilized on a score by score basis. The Pearson r and Spearman r were used to estimate the correlation between a BAT selection composite based on first and second administration scores. The BAT selection composite combined 17 BAT summary scores using the beta weights from the Pilot Selection and Classification System (PSACS) pilot candidate selection model (but without the AFOQT scores). For the tests involving psychomotor skills (Two-Hand Coordination, Complex Coordination, and Time-Sharing 3), the reliability estimates were for tracking error/difficulty "scores" summed over time. For the other tests, the "scores" refer to test items.

The Guttman split-half reliability coefficient (Guttman, 1945) is coefficient alpha applied to a two-item scale (i.e., to the sum of scores across items). In this study, the day 1 sum of scores and the day 2 sum of scores for a particular test attribute make up the two-item scale for that test attribute. Due to the relaxed assumptions for this coefficient, the true reliability of the scale tends to be underestimated.

III. RESULTS

Descriptive Statistics

Tables 2 and 3 provide descriptive statistics for subjects' performance on the first and second BAT battery administration. Score distributions for many of the BAT scores are non-normal. Tracking error/tracking difficulty scores and response time scores tend to be strongly skewed and overrepresented at the extremes. As this is common with scores of this type, no effort was made to normalize the distributions through score transformations, recoding, or removal of extreme scores.

Table 2. **Descriptive Statistics for the Basic Attributes Test Battery: First Administration**

Test Score	Mean	SD	Minimum	Maximum	Skew	Kurtosis
Two-Hand Coordination						
X-Axis Tracking Error - (horizontal)	5,023.9	1,967.6	2,394.0	24,966.0	5.2	48.7
Y-Axis Tracking Error - (vertical)	5,903.8	1,713.7	2,827.0	15,671.0	1.7	5.6
Complex Coordination						
X-Axis Tracking Error - (horizontal)	14,253.2	10,481.5	621.0	57,369.0	1.4	1.7
Y-Axis Tracking Error - (vertical)	9,165.8	7,643.6	995.0	71,942.0	3.3	19.6
Z-Axis Tracking Error - (rudder bar)	8,081.0	6,213.6	1,296.0	56,248.0	3.1	16.6
Encoding Speed 2						
Avg RT (ms) - Correct responses	767.3	124.4	541.6	1,357.5	1.3	2.7
Percent Correct (%)	91.6	4.9	68.8	100.0	-1.0	1.8
Mental Rotation 2						
Avg RT (ms) - Correct responses	922.6	260.7	455.5	1,789.1	0.9	0.3
Percent Correct (%)	90.8	9.1	47.9	100.0	-2.3	6.2
Item Recognition 2						
Avg RT (ms) - Correct responses	749.0	189.7	457.7	1,704.0	1.6	3.9
Percent Correct (%)	93.4	6.0	70.8	100.0	-1.0	0.7
Time-Sharing 3						
Avg Tracking Difficulty - Single Task	310.0	73.1	19.0	501.0	-0.5	1.9
Avg Tracking Difficulty - Dual task	240.2	34.9	32.3	308.8	-0.8	3.0
Avg RT (ms) - Dual task	558.4	111.4	414.2	1,572.8	4.6	33.7
Self-Crediting Word Knowledge						
Avg RT (ms) - Correct responses	6,586.9	1,508.7	3,480.8	11,920.2	0.3	0.0
Percent Correct (%)	63.9	9.8	40.0	86.7	0.0	-0.3
Avg Bet	38.9	5.2	10.0	50.0	-0.4	-0.2
Activities Interest Inventory 2						
Avg RT (ms) - all responses	3,938.9	952.2	1,531.4	7,246.3	0.5	0.2
Number of High-Risk Choices	29.0	5.4	3.0	41.0	-0.6	1.2

Notes.

1. Skewness is a statistic needed to determine the degree to which a distribution of scores approximates a normal curve, as it measures deviations from symmetry. It will equal zero when the distribution is a completely symmetric, bell-shaped curve. Positive values indicate that most of the scores are clustered below the mean, with most extreme values to the right. Negative values indicate clustering above the mean, with most extreme values to the left.

2. Kurtosis is a measure of the relative peakedness or flatness of the curve defined by the distribution of scores. A normal distribution will have a kurtosis of zero. Positive values indicate greater peakness, whereas negative values indicate a flatter than normal distribution.

N = 247

Table 3. **Descriptive Statistics for the Basic Attributes Test Battery: Second Administration**

Test Score	Mean	SD	Minimum	Maximum	Skew	Kurtosis
Two-Hand Coordination						
X-Axis Tracking Error - (horizontal)	3,747.1	1,321.5	2,103.0	15,549.0	4.9	36.3
Y-Axis Tracking Error - (vertical)	4,047.2	1,273.7	2,217.0	15,035.0	4.2	28.6
Complex Coordination						
X-Axis Tracking Error - (horizontal)	7,527.0	7,418.3	350.0	45,985.0	2.4	7.4
Y-Axis Tracking Error - (vertical)	5,701.6	7,224.6	414.0	70,236.0	4.9	32.7
Z-Axis Tracking Error - (rudder bar)	4,956.1	5,624.6	757.0	72,000.0	7.6	83.0
Encoding Speed 2						
Avg RT (ms) - Correct responses	667.6	134.2	404.6	1,152.5	1.3	2.1
Percent Correct (%)	91.6	6.3	54.2	100.0	-1.5	5.1
Mental Rotation 2						
Avg RT (ms) -	722.8	193.6	324.1	1,420.6	1.1	1.5
Percent Correct (%)	90.3	8.0	56.3	100.0	-1.7	3.7
Item Recognition 2						
Avg RT (ms) - Correct response	681.1	176.4	277.5	1,598.0	1.3	3.9
Percent Correct (%)	92.8	6.9	62.5	100.0	-1.4	2.6
Time-Sharing 3						
Avg Tracking Difficulty - Single task	349.5	73.8	153.2	607.6	-0.1	0.6
Avg Tracking Difficulty - Dual task	234.4	35.8	129.1	314.0	-0.5	0.4
Avg RT (ms) - dual task	509.8	83.5	377.8	956.6	2.1	7.2
Self-Crediting Word Knowledge						
Avg RT (ms)- Correct responses	4,065.6	1,185.7	2,109.6	11,920.0	2.2	8.8
Percent Correct (%)	64.0	9.3	33.3	90.0	-0.1	0.4
Avg Bet	36.6	9.2	10.0	50.0	0.0	-1.0
Activities Interest Inventory 2						
Avg RT (ms) - All responses	2,571.1	758.5	296.0	6,026.4	1.3	4.1
Number of High-Risk Choices	29.5	5.3	13.0	41.0	-0.5	-0.2

Notes.

1. Skewness is a statistic needed to determine the degree to which a distribution of scores approximates a normal curve, as it measures deviations from symmetry. It will equal zero when the distribution is a completely symmnetric, bell-shaped curve. Positive values indicate that most of the scores are clustered below the mean, with most extreme values to the right. Negative values indicate clustering above the mean, with most extreme values to the left.

2. Kurtosis is a measure of the relative peakedness or flatness of the curve defined by the distribution of scores. A normal distribution will have a kurtosis of zero. Positive values indicate greater peakness, whereas negative values indicate a flatter than normal distribution.

N = 247

Comparison of the means in Table 2 with those in Table 3 indicate general improvement in tracking and response time performance on the second test administration (i.e., lower tracking error on Two-Hand Coordination and Complex Coordination; higher tracking difficulty on Time-Sharing 3; and quicker response times on Encoding Speed 2, Mental Rotation 2, Item Recognition 2, Time-Sharing 3, Self-Crediting Word Knowledge, and Activities Interest Inventory 2). The psychomotor test results are consistent with those from Mercatante (1988), who reported a significant improvement in tracking performance on the Two-Hand Coordination and Complex Coordination tests after a 2- to 3-week test-retest interval. Although average response time became quicker, response accuracy changed very little. Table 4 summarizes the results of paired t-tests comparing individual BAT scores and a BAT composite for the first and second administrations.

Table 4. **Paired T-Tests Comparing First and Second Administration Test Scores**

Test Score	Mean Admin 1	Mean Admin 2	Paired T-Test Value
Two-Hand Coordination			
X-Axis Tracking Error (horizontal)	5,023.9	3,747.1	10.5*
Y-Axis Tracking Error (vertical)	5,903.8	4,047.2	13.4*
Complex Coordination			
X-Axis Tracking Error (horizontal)	14,253.2	7,527.0	11.2*
Y-Axis Tracking Error (vertical)	9,165.8	5,701.6	6.4*
Z-Axis Tracking Error (rudder)	8,081.0	4,956.1	6.8*
Encoding Speed 2			
Avg RT (ms) - correct responses	767.3	667.6	13.9*
Percent Correct (%)	91.6	91.6	-0.1
Mental Rotation 2			
Avg RT (ms.) - correct responses	922.6	722.8	16.5*
Percent Correct (%)	90.8	90.3	0.9
Item Recognition 2			
Avg RT (ms.) - correct responses	749.0	681.5	8.9*
Percent Correct (%)	93.9	92.8	1.2
Time-Sharing 3			
Avg Tracking Difficulty - Single task	310.0	349.5	-7.7*
Avg. Tracking Difficulty - Dual task	240.2	234.4	1.4
Avg RT (ms) - Dual Task	558.4	509.8	7.7*
Self-Crediting Word Knowledge			
Avg RT (ms) - Correct responses	6,586.9	4,065.6	29.4*
Percent Correct (%)	63.9	64.0	-0.2
Avg Bet	38.9	36.6	1.6
Activities Interest Inventory			
Avg RT (ms) - all responses	3,938.9	2,571.1	25.7*
Number of High-Risk Choices	29.0	29.5	-1.8
BAT Composite	50.0	60.5	-7.5*

Note. The t-test probability is for a two-tailed test.
$N = 247$
$^{*}p \leq .01$

Reliability

Table 5 summarizes the internal consistency reliability estimates for the shortened BAT battery used in this study and compares them to estimates from full-length tests. In general, the test scores from this study demonstrated acceptable internal consistency for scores of this type. Cronbach's alpha ranged between .92 and .97 for tracking error (Two-Hand and Complex Coordination) and tracking difficulty (Time-Sharing 3) scores. Scores based on response times also showed good internal consistency (Cronbach's alpha between .88 and .97). Internal consistency estimates were lower for response accuracy (i.e., correct or incorrect: Encoding Speed 2, Item Recognition 2, Mental Rotation 2, and Self-Crediting Word Knowledge; Cronbach's Alpha between .30 and .84) and response choice (Activities Interest Inventory 2; Cronbach's Alpha between .77 and .80) measures. This is not surprising, as internal consistency tends to be low for test items with only two response alternatives (i.e., yes/no or same/different) and for low difficulty levels where response speed is the major discriminating factor among individuals. Reducing the length of some tests did not seem to lower the internal consistency of several of the test scores (Carretta, 1990a; response time - Encoding Speed, Mental Rotation, Item Recognition, and Activities Interest Inventory; tracking difficulty - Time-Sharing; and response choice - Activities Interest Inventory). The greatest reduction in reliability occurred for response accuracy scores, which typically have low to moderate reliability on the full-length tests.

As shown in Table 6, only 2 of 16 correlations between forms and 8 of 16 Guttman split-half reliability estimates exceeded .70. This suggests that even though performance generally improved on the second administration, individual subjects changed by differing amounts (possibly due to ceiling and floor effects or regression toward the mean).

The BAT composite score demonstrated a modest correlation between the two administrations (Pearson r = .56; Spearman r = .55). A simple sign test showed that 172 of the 247 subjects (69.6%) improved their BAT composite score on the second administration (z = 6.2, p < .01). There is some evidence of regression toward the mean on the second administration. Only 66 of 124 (53.2%) subjects with BAT composite scores above the median on the first administration improved their scores on the second administration whereas 106 of 123 (86.6%) subjects with a composite score below the median on the first administration improved their score on the second administration. Table 7 summarizes BAT composite score gains and losses grouped by 20% groups (i.e., quintiles) on the first administration; Figure 1 illustrates average BAT percentile rank score gains and losses grouped by 20% groups (top, 2nd, 3rd, 4th and bottom 20% groups) on the first administration. On average, subjects improved their BAT percentile ranking by 10.5 points. The bottom three 20% groups on the first administration improved their percentile ranking on the second administration, whereas those who scored in the top two 20% groups on the first administration performed worse on the second administration.

IV. DISCUSSION

Test-retest reliability estimates for individual BAT summary scores were somewhat stronger than those reported for two separate studies with computerized cognitive abilities test batteries (Roznowski, 1989; Saccuzzo & Larson, 1987) and for a US Air Force psychomotor coordination test battery (Mercatante, 1988). This would be expected as the test-retest interval was 10 days in the Saccuzzo and Larson (1987) study, 2 weeks in the Roznowski (1989) study, and 2 to 3 weeks in the Mercatante (1988) study, but only 1 day in the present study. Results from these analyses suggest an upper limit of agreement of about .56 between a BAT selection composite on the experimental BAT system and one generated from scores from an operational test system.

Table 5. Reliability Estimates for the Basic Attributes Test Battery: Shortened Versus Full-Length Tests

Test Score	Current Study			Full-Length Tests	
	Number of Scores per Admin	Cronbach's Alpha Admin.1	Cronbach's Alpha Admin.2	Number Of Scores	Cronbach's Alpha
Two-Hand Coordination					
X-Axis Tracking Error - (horizontal)	10	.94	.93	10	.94
Y-Axis Tracking Error - (vertical)	10	.92	.92	10	.95
Complex Coordination					
X-Axis Tracking Error - (horizontal)	10	.95	.95	10	.95
Y-Axis Tracking Error - (vertical)	10	.95	.97	10	.99
Z-Axis Tracking Error	10	.94	.97	10	.94
Encoding Speed 2					
Response Time per Trial	48	.94	.94	96	.96
Response Outcome per Trial - (Correct/incorrect)	48	.41	.68	96	.71
Mental Rotation 2					
Response Time per Trial	48	.96	.97	72	.97
Response Outcome per Trial - (Correct/incorrect)	48	.84	.80	72	.90
Item Recognition 2					
Response Time per Trial	24	.89	.88	48	.95
Response Outcome per Trial - (Correct/incorrect)	24	.30	.47	48	.54
Time Sharing 3					
Tracking Difficulty	11	.96	.96	19	.96
Self-Crediting Word Knowledge					
Response Time per Trial	30	.89	.94	30	.89
Response Outcome per Trial - (Correct/incorrect)	30	.55	.49	30	.65
Activities Interest Inventory 2					
Response Time per Trial	41	.92	.95	81	.95
Response Choice per Trial - (Low/high risk)	41	.77	.80	81	.86

Notes.

1. The column labeled "Number of Scores per Administration" refers to summed tracking error/tracking difficulty measures for the Two-Hand Coordination, Complex Coordination, and Time-Sharing 3 tests. For all other tests, "Number of Scores per Administration" refers to the number of test items.

2. Full-length test reliability estimates were reported previously in Carretta (1990a).

N = 247

Table 6. **Test-Retest Reliability Estimates for the Basic Attributes Test Battery**

Test Score	Number of scores per admin	Corr. between forms	Guttman split-half
Two-Hand Coordination			
X-Axis Tracking Error - (horizontal)	10	.46	.58
Y-Axis Tracking Error - (vertical)	10	.54	.65
Complex Coordination			
X-Axis Tracking Error - (horizontal)	10	.49	.62
Y-Axis Tracking Error - (vertical)	10	.40	.56
Z-Axis Tracking Error	10	.26	.41
Encoding Speed 2			
Response Time per Trial	48	.50	.65
Response Outcome per Trial - (Correct/incorrect)	48	.26	.40
Mental Rotation 2			
Response Time per Trial	48	.68	.79
Response Outcome per Trial - (Correct/incorrect)	48	.55	.71
Item Recognition 2			
Response Time per Trial	24	.66	.79
Response Outcome per Trial - (Correct/incorrect)	24	.39	.55
Time-Sharing 3			
Tracking Difficulty	11	.67	.80
Self-Crediting Word Knowledge			
Response Time per Trial	30	.56	.72
Response Outcome per Trial - (Correct/incorrect)	30	.71	.86
Activities Interest Inventory 2			
Response Time per Trial	41	.54	.70
Response Choice per Trial - (Low/high risk)	41	.75	.86

Note. The column labeled 'Number of Scores per Administration' refers to summed tracking error/tracking difficulty measures for the Two-Hand Coordination, Complex Coordination, and Time-Sharing 3 tests. For all other tests, 'Number of Scores per Administration' refers to the number of test items.

N = 247

Table 7. Number of Subjects Who Showed Improvement or Decrement on BAT Composite at Second Administration

First Admin. Quintile	N	Improvement	Decrement
1 (top)	50	20	30
2	49	28	21
3	49	39	10
4	49	40	09
5 (bottom)	50	45	05
TOTAL	247	172	75

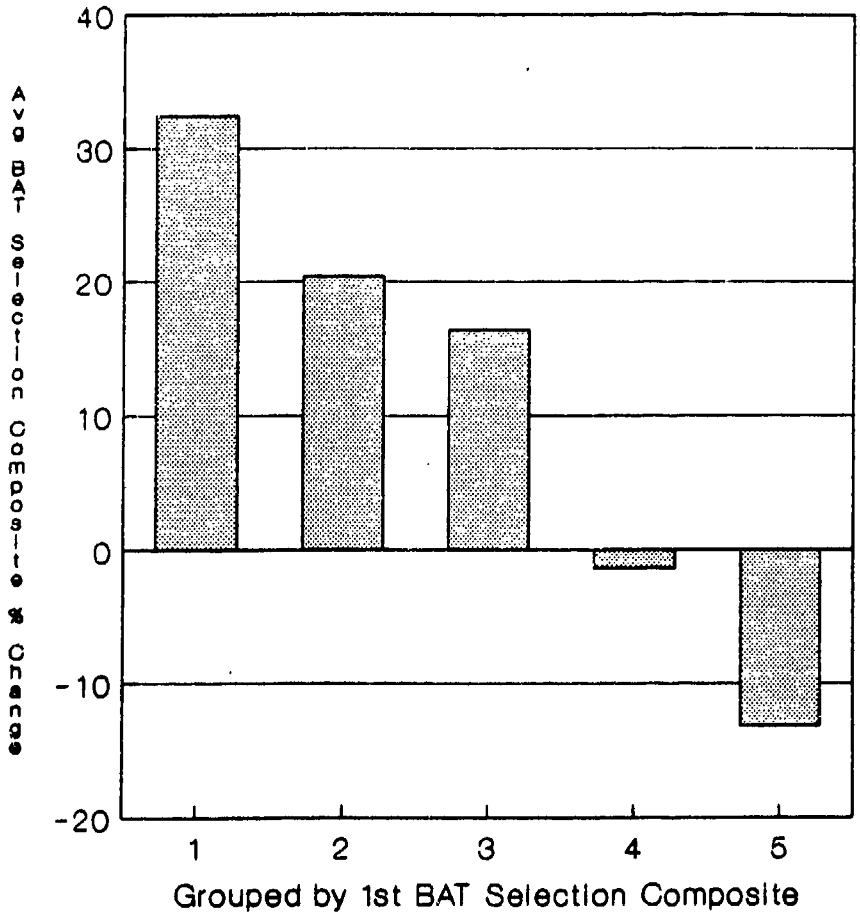

Figure 1. BAT Pilot Selection Composite Gains/Losses on 2nd Administration.

The BAT composite test-retest estimate may be low for several reasons. First, for administrative reasons it was decided to use shortened versions of five of the eight BAT battery tests in the present study. If a full-length battery had been used, approximately 25% fewer subjects would have been able to participate in the study. As a trade-off, the reliability of the shortened tests and that of the BAT composite were reduced somewhat. As noted earlier, the internal consistency of test scores based on response times or tracking difficulty was not reduced significantly by shortening these tests. However, the shortened tests did show lower internal consistency for the response accuracy scores. Second, the composition of the sample used in this study may have reduced the strength of the reliability estimates, due to range restrictions on the abilities measured by the AFOQT. The ROTC students in this study had already been selected for commissioning in the USAF and for pilot training, in part on the basis of their AFOQT scores and college performance. Without testing ROTC applicants on the BAT battery, possible restriction of range on BAT scores due to ROTC selection procedures for pilot training cannot be estimated. Finally, individual differences in intrinsic motivation may have affected performance on the test battery and, as a result, decreased its test-retest reliability. The subjects who participated in this study had no extrinsic motivation to perform well or to try to do better on the second test administration. One possible interpretation of the BAT composite change scores is that those who performed poorly on the first administration tried harder to improve their performance on the second administration.

V. CONCLUSION

Despite administrative, sampling, and motivational circumstances that may have reduced the subtest and composite score reliability estimates, agreement between the time 1 - time 2 BAT pilot selection composite scores was moderate. Some improvement in the test-retest correlation can be expected if scores with poor reliability are removed from the integrated pilot candidate selection composite. The trade-off here could be a reduction in validity.

Results from this study suggest that short-term retesting on the BAT battery should be prohibitied except when administrative (e.g., BAT system failure) or medical (e.g., illness during testing) problems justify a short-term retest. When a short-term retest is allowed, it may be necessary to adjust the scores on the retest to minimize test score gains due to prior exposure to the test. Additional studies need to be conducted to determine the length of time needed to avoid retest gains due to prior exposure to the BAT battery. Mercatante (1988) reported a significant improvement in tracking performance on the Two-Hand Coordination and Complex Coordination tests after a 2- to 3-week test-retest interval.

The study to evaluate the equivalency between the experimental and operational BAT systems can avoid the problems encountered here by using a full-length BAT battery and constructing the sample such that it reflects the USAF pilot training applicant population. That is, subjects' scores should reflect the full range of performance that can be expected from USAF pilot training applicants. Ideally, subjects should be sampled from pilot training applicants from each commissioning source in direct proportion to the number of pilot candidates commissioned through that source. If this is not feasible, other types of subjects may be substituted (e.g., basic recruits, college students) to the extent that they can be shown to be similar to USAF pilot training applicants.

REFERENCES

Bordelon, V.P., & Kantor, J.E. (1986). *Utilization of psychomotor screening for USAF pilot candidates: Independent and integrated selection methodologies* (AFHRL-TR-86-4, AD-A170 353). Brooks AFB TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Carretta, T.R. (1989). USAF pilot selection and classification systems. *Aviation Space and Enviromental Medicine, 60*(1), 46-49.

Carretta, T.R. (1990a). *Basic Attributes Test (BAT): A preliminary comparison between Reserve Officer Training Corps (ROTC) and Officer Training School (OTS) pilot candidates* (AFHRL-TR-89-50, AD-A224 093). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Carretta, T.R. (1990b). *Cross-validation of experimental USAF pilot training performance models* (AFHRL-TR-89-68, AD-A222 253). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

Cronbach, L.J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrica, 10*(4), 255-282.

Kantor, J.E., & Carretta, T.R. (1988). Aircrew selection systems. *Aviation Space and Environmental Medicine 59*(11) Supplement, A32-A38.

Mercatante, T.A. (1988). *The reliability and validitiy of psychomotor aptitude for pilot selection.* Unpublished master's thesis, St. Mary's University, San Antonio, TX.

Passey, G.E., & McLaurin, W.A. (1966). *Perceptual-psychomotor tests in aircrew selection: Historical review and advanced concepts* (PRL-TR-66-4, AD-636 606). Lackland AFB, TX: Personnel Research Laboratory.

Roznowski, M. (1989). Evidence for the construct validity of experimental cognitive tasks. *Proceedings of the 31st Annual Conference of the Military Testing Association*, 173-178.

Saccuzzo, D.P., & Larson, G.E. (1987). *Analysis of test-retest reliability of cognitive speed tests* (NPRDC TR-88-10). San Diego, CA: Navy Personnel Research and Development Center.

Skinner, J., & Ree, M.J. (1987). *Air Force Officer Qualifying Test (AFOQT): Item and factor analyses of Form O* (AFHRL-TR-86-68, AD-A184 975). Brooks AFB, TX: Manpower and Personnel Division, Air Force Human Resources Laboratory.

United States Air Force (1983). *Application procedures for UPT, UPTH and UNT*, Air Force Regulation 51-4. Washington, DC: Department of the Air Force.

APPENDIX: DESCRIPTION OF THE BASIC ATTRIBUTES TEST BATTERY

BAT BATTERY DESCRIPTIONS

Test Battery Introduction

This interactive subprogram prompts the subject to provide background information (e.g., identity, age, gender), as well as personal history and attitudes related to flying.

Two-Hand Coordination

The Two-Hand Coordination test is a variation of a rotary pursuit task. The airplane (target) moves in an eliptical path on the screen at a rate of 20 cycles per minute. The rate of movement of the airplane within each cycle varies in a fixed sinusoidal pattern. The subject controls the vertical and horizontal movement of a small "gunsight" using a left and right joystick. The left-hand joystick controls the vertical movement of the gunsight; the right-hand joystick controls the horizontal movement of the gunsight. The subject's task is to keep the gunsight centered on the moving airplane. After receiving instructions, the subject completes a 3-minute practice session and a 5-minute test. The measures of interest are horizontal and vertical tracking error scores and axis stick movement rate scores. The psychological factors for this test are low- to moderate-order tracking and time-sharing ability in pursuit.

Complex Coordination

Complex Coordination uses a dual-axis joystick (right-hand) to control the horizontal and vertical movement of a cursor. The left-hand joystick controls the left-right movement of a vertical "rudder bar" of light at the base of the screen. The subject's task is to center the cursor (against a constant horizontal and vertical rate bias) on a large cross fixed at the center of the screen while simultaneously centering the rudder bar at the base of the screen (also against a constant rate bias). The instruction, practice, testing, and scoring are as in the Two-Hand Coordination test. The Complex Coordination test assesses compensatory tracking ability involving multiple axes. This test requires about 10 minutes to complete.

Encoding Speed 2

The subject is presented simultaneously with two letters and is required to make a same/different judgment about the letter pair. According to three different conditions, the judgment may be based on physical identity (AA vs. Aa), name identity (AA vs. AH), or category identity (vowels vs. consonants - AE vs. AH). The latency of the encoding judgment provides a measure of the speed of the cognitive encoding process.

Reaction time and accuracy of response (correct/incorrect) are recorded on each of the 48 trials (16 trials in each condition) of this shortened test. The full-length version has 96 trials and requires about 20 minutes to complete. The psychological factor involved in this test is verbal classification at several levels of cognitive operation.

Mental Rotation 2

The subject is presented sequentially with a pair of letters and is required to make a same/different judgment. The letter pair may be either identical or mirror images, and the pair may be either in the same orientation or rotated in space with respect to each other. A correct "different" judgment is associated with a mirror image pair and is not dependent on the relative rotation of the two letters.

In order to perform the task, the subject must form a mental image of the first letter (no longer displayed) and perform a point-by-point comparison with the second letter (which remains on the screen). In addition, when the letters are rotated with respect to each other, the subject must mentally rotate the mental image of one letter into congruence with the other prior to making the comparison.

Speed and accuracy of response are recorded on each of the 48 trials of this shortened test. The full-length version has 72 trials and requires about 25 minutes to complete. The psychological factor assessed by this test is spatial transformation ability.

Item Recognition 2

In this test, a string of one to six digits is presented on the screen. The string is then removed and followed, after a brief delay, by a single digit. The subject is instructed to remember the initial string of digits, then to decide if the single digit was one of those presented in the initial string. The subject is instructed to respond by pressing a keypad button marked YES if the single digit was in the string or another marked NO if it was not. The instructions inform the subject to work as quickly and accurately as possible. Speed and accuracy of response are recorded on each of the 24 trials of this shortened test. There are 48 trials on the full-length version of this test, which requires about 20 minutes to complete. Short-term memory storage, search, and comparison operations are the underlying psychological factors for this test.

Time-Sharing 3

This shortened version of the Time-Sharing test has 11 1-minute trials (five practice, tracking only trials and six 1-minute, dual-task trials). The normal-length version has 19 1-minute trials (see below).

During a series of five 1-minute trials, the subject is required to learn a compensatory tracking task. To perform this task, the subject must anticipate the movement of a marker on a screen and operate a control stick to counteract that movement in order to keep the marker aligned with a fixed central point. Throughout the task, task difficulty is adjusted based on the subject's performance. The control dynamics are a combination of rate and acceleration components. The "disturbance" factor is a quasi-random, summed sinusoidal forcing function.

After five 1-minute "tracking only" trials, the subject is required to track while cancelling digits that appear at random intervals and locations on the screen (six 1-minute trials). The subjects cancels a digit by pressing a like-numbered button on the response keypad. A "cross-adaptive" logic forces the subject to respond to digits within 4 seconds after their appearance. If the subject fails to respond within 4 seconds, he/she loses control of the gunsight until a correct response is made. These dual-task trials occur in two 3-minute blocks. The information processing load gradually increases during these trials. The full-length version

of this test has 19 1-minute trials as follows: 10 1-minute "tracking only" trials; six 1-minute dual task trials, and three more 1-minute "tracking only" trials.

The effects of the secondary task loads are reflected in the pattern of level-of-difficulty changes caused by the adaptive logic that holds tracking error constant. The measure of interest for this test is the level of difficulty at which the subject can perform consistently. This test assesses a variety of psychological factors including higher-order tracking ability, learning rate, and time sharing ability as a function of differential task load.

Self-Crediting Word Knowledge

This test is essentially a vocabulary test in which the subject is presented with a "target" word and five other words from which the closest synonym must be chosen. There are three blocks of 10 questions each. The target words become increasingly difficult with each successive block. The subject is informed of this increasing difficulty and is required to make a bet prior to each block as to how well the subject expects to do. Response time and accuracy are recorded on each of the 30 trials, which require about 10 minutes to complete. This test assesses self-assessment ability and self-confidence.

Activities Interest Inventory 2

This shortened test is designed to determine the subject's interest in various activities. The subject is presented with 41 pairs of activities and is asked to choose between them. (There are 81 pairs in the full-length version of this test.) The subject is told to assume that he/she has the necessary ability to perform each activity. The activity pairs force the subject to choose between tasks that differ on threat to physical survival--sometimes subtly, sometimes not. The measures of interest are the number of high-risk options chosen and the amount of time required to choose between pairs of activities. The psychological factor assessed by this test is attitudes toward risk. The 81-item, full-length version of this test requires about 10 minutes to complete.